Mapping design values over Canada

Chao Li et al.

March 5, 2018

Data formatting

We represent the inferred design values (e.g., 100-year extreme rainfall) in a region from the large ensemble CanRCM4 simulations by an $n \times p$ data matrix $\boldsymbol{X}$, with n being the ensemble size of the simulations and p the number of grid cells in the region. The region can be the entire land of Canada or a subregion. We center the data matrix by subtracting the ensemble mean of the design values at each grid cell, that is, $\boldsymbol{X}^{'} = \left(\boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n\right)\boldsymbol{X}$, where $\boldsymbol{I}_n$ is the identity matrix of size n and $\mathbf{1}_n$ is the $n \times n$ all-ones matrix. To minimize the influence of the varying grid cell areas on the structure of the computed EOFs, we weight each value in $\boldsymbol{X}^{'}$ by the grid cell area, that is, $\boldsymbol{X}^{'}_{w} = \boldsymbol{X}^{'}\boldsymbol{W}$, with $\boldsymbol{W} = \mathrm{diag}\left(f_1, f_2..., f_p\right)$ being the weighting matrix and f_i the fractional area of the grid cell i. In the following and unless otherwise stated, we drop the super- and subscripts and denote the centered and weighted design value matrix by $\boldsymbol{X}$. The spatial covariance of $\boldsymbol{X}$ is estimated by $\boldsymbol{C} = \boldsymbol{X}^T\boldsymbol{X}$.

EOF with data matrix

EOF analysis can be computed by singular value decomposition (SVD) of the data matrix $\boldsymbol{X}$, or by eigendecomposition of the covariance matrix $\boldsymbol{C}$. According

to SVD, the $n \times p$ matrix $\boldsymbol{X}$ can be factored as

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T \tag{1}$$

where $\boldsymbol{U}$ is an $n \times n$ orthogonal matrix whose columns $\boldsymbol{u}_i$ $(i = 1, 2, ..., n)$ are the EOF coefficients which are often as well referred to as principal components (PCs); $\boldsymbol{V}$ is a $p \times p$ orthogonal matrix whose columns $\boldsymbol{v}_i$ $(i = 1, 2, ..., p)$ provides the EOFs; and $\boldsymbol{\Sigma}$ is an $n \times p$ diagonal matrix of the form diag $(\sigma_1, \sigma_2, ..., \sigma_r, 0, ..., 0)$ with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ and $r = \text{rank}(\boldsymbol{X}) \leq \min(n, p)$. In the above, σ_1^2, ..., σ_r^2 are the eigenvalues of $\boldsymbol{X}^T\boldsymbol{X}$ (also of the covariance matrix $\boldsymbol{C}$).

EOF with covariance matrix

The covariance matrix $\boldsymbol{C}$ (of size $p \times p$) has an eigendecomposition in the form of

$$\boldsymbol{C} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T \tag{2}$$

where $\boldsymbol{V}$ is a $p \times p$ orthogonal matrix containing the EOFs $\boldsymbol{v}_i$ $(i = 1, 2, ..., p)$ in its columns; $\boldsymbol{\Lambda}$ is a $p \times p$ diagonal matrix containing the eigenvalues of $\boldsymbol{C}$ down the diagonal, that is, $\boldsymbol{\Lambda} = \text{diag}(\sigma_1^2, \sigma_2^2, ..., \sigma_r^2, 0, ..., 0)$ with $\sigma_1^2 \geq \sigma_2^2 \geq \cdots \geq \sigma_r^2 > 0$ and $r \leq \min(n, p)$. Projecting $\boldsymbol{X}$ onto the EOFs by $\boldsymbol{A} = \boldsymbol{X}\boldsymbol{V}$ obtains, in the columns of $\boldsymbol{A}$, the EOF coefficients. Unlike those computed via SVD, here the EOF coefficients $\boldsymbol{\alpha}_i$ $(i = 1, 2, ..., p)$ are not necessarily unitary. With the EOFs and their coefficients, the data matrix $\boldsymbol{X}$ can be factored as

$$\boldsymbol{X} = \boldsymbol{A}\boldsymbol{V}^{-1} = \boldsymbol{A}\boldsymbol{V}^T \tag{3}$$

EOF reconstruction: the basis of the mapping approach

The EOFs describe the spatial dependence structure of the design value field, while the EOF coefficients reflect the variation of the EOF modes due to internal variability. The eigenvalues give a measure of the importance of each EOF mode. Only the first r EOF modes with nonzero eigenvalues carry information. Neglecting the redundant ones, the data matrix $\boldsymbol{X}$ can be rewritten in vector form from (1) as follows

$$\boldsymbol{X} = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T \tag{4}$$

Likewise, from (3), we have

$$\boldsymbol{X} = \sum_{i=1}^{r} \boldsymbol{\alpha}_i \boldsymbol{v}_i^T \tag{5}$$

The reconstructions (4-5) state that the design value field can be fully represented by a set of basis functions of space (i.e., EOFs) and a set of amplitude functions of internal variability (i.e., EOF coefficients). This provides the foundamental of our approach to design value mapping.

In fact, it is a few leading EOFs that are capturing large-scale and low-frequency features of the field, with the remaining mostly representing random noise. The reconstruction is therefore typically implemented by truncating the above sums at some m ($\ll r$) such that only the important large-scale features are retained while the random noise is filtered. In practice, the truncation level m is determined by prescribing a certain amount of the explained variance (e.g., 95%) and choosing the first m EOFs that explain altogether this amount of variance. The explained variance by the first m EOFs is computed as a percent by

$$\frac{100 \sum_{i=1}^{m} \sigma_i^2}{\sum_{i=1}^{r} \sigma_i^2} \%$$

The design value mapping method

Given the relatively sparse distribution of meteorological observing stations in Canada plus the fact that many of these stations cover too short periods to allow reliable estimates of design values, traditional interpolation methods such as spline interpolcation and Kriging become impractical. We thus base the mapping method on EOF reconstruction. To be specific, the method obtains leading EOF modes of design values from a 50-member ensemble of CanRCM4 simulations, constrains the amplitudes of these modes (i.e., EOF coefficients) by the available observed design values via a least squares regression, and then combines them together to reconstruct design values at unobserved locations. The underlying assumption is that CanRCM4 simulations can reasonably resolve the overall spatial structure of observed design value field, despite the possibility that the magnitudes of design values might differ between models and observations.

The mapping method needs a dataset of observed design values gridded consistently with CanRCM4 simulations. For that, we grid the design values estimated from observations at stations with sufficient data by averaging available values within each CanRCM4 grid cell ($\sim$ 50 km $\times$ 50 km). Grid cells with no observations are flagged as missing. The gridded observations are then organized into a matrix $\boldsymbol{y}_g$ (which is actually a $1 \times p$ vector) and weighted by grid cell area, both following the way as for CanRCM4 simulations. We use the subscript "g" to make it explicit that the matrix contain gaps (i.e., missing values). For operational convenience, we treat $\boldsymbol{y}_g$ as a *column* vector in the following.

With $\boldsymbol{X}$ and $\boldsymbol{y}_g$, the maping method is implemented as follows:

1) Calculate EOFs $\boldsymbol{V}$ and eigenvalues, by SVD of $\boldsymbol{X}$ or eigendecomposition of $\boldsymbol{X}^T\boldsymbol{X}$.

2) For a given number k of EOFs, find the observation-constrained EOF coefficients $\hat{\boldsymbol{\beta}} = \left(\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_k\right)^T$ via a least sequares regression, $\boldsymbol{y}_{g,\perp} = \boldsymbol{V}_{1:k,\perp}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where subscript $1:k$ means that the first k columns of $\boldsymbol{V}$ (which correspond to the first k EOFs) are involved in the regression; and the subscript $\perp$ indicates collapsed versions of the corresponding matrices with rows that do not have observations removed. The solution $\hat{\boldsymbol{\beta}}$ is found by minimizing the objective function $\psi_k = || \boldsymbol{y}_{g,\perp} - \boldsymbol{V}_{1:k,\perp}\boldsymbol{\beta} ||^2$, giving $\hat{\boldsymbol{\beta}} = \left(\boldsymbol{V}_{1:k,\perp}^T \boldsymbol{V}_{1:k,\perp}\right)^{-1} \boldsymbol{V}_{1:k,\perp}^T \boldsymbol{y}_{g,\perp}$.

3) Repeat 2) by increasing k to $k+1$ until the change of the objective function is $< 10^{-3}$ or the explained variance by the involved EOFs is $\geq 95\%$. If neither of the criteria is fullfilled but the number of involved EOFs becomes greater than 10% of the number of the grid cells having observations, the method is marked as not converged, otherwise denote the convergence step as l and proceed to 4).

4) Reconstruct the full design value field and remove the effect of areal weighting from the reconstructions by conducting $\boldsymbol{y} = \boldsymbol{V}_{1:l}\hat{\boldsymbol{\beta}}\boldsymbol{W}^{-1}$. In order to respect grid cells with observations, relevant reconstructed values are substituted by the corresponding observations. The reconstructions are then used to plot a map for design values with full coverage over Canada.